JOURNAL OF APPLIED BEHAVIOR ANALYSIS 1992, 25, 101–116 NUMBER 1 (SPRING 1992)

THE CLASSWIDE PEER TUTORING PROGRAM: IMPLEMENTATION FACTORS MODERATING STUDENTS' ACHIEVEMENT

CHARLES R. GREENWOOD, BARBARA TERRY, CARMEN ARREAGA-MAYER, AND REBECCA FINNEY

JUNIPER GARDENS CHILDREN'S PROJECT, UNIVERSITY OF KANSAS

We conducted a study designed to assess implementation of the classwide peer tutoring program and the relationship between implementation variation and student outcome. A clinical replication design was used. Five volunteer elementary teachers were trained to implement the program; their implementation was monitored for 19 consecutive weeks during 1 school year. Overall, the results indicated that specific variations in program implementation were associated with students' responses to treatment. It was also demonstrated that different teachers' applications of the program produced differential levels of student outcome. Implementation factors related to lower spelling achievement were (a) reduced opportunities to receive program sessions, (b) reduced probabilities of students' participation in program opportunities, (c) too many students assigned unchallenging spelling words, and (d) reduced rates of daily point earning reflecting lower levels of spelling practice during tutoring sessions. The implications of these findings and methods of preventing these implementation problems are discussed in the context of quality assurance and social validity.

DESCRIPTORS: peer tutoring, education, fidelity, academic behavior, assessment

The relationship between treatment effectiveness and the strength and fidelity of treatment implementation has become an important issue in recent years (Carta & Greenwood, 1989; Peterson, Homer, & Wonderlich, 1982; Yeaton & Sechrest, 1981). Strength of treatment has been described as the intensity of the treatment agent (e.g., drug dosage, duration of treatment), whereas fidelity of treatment reflects the extent to which a treatment protocol is accurately implemented (Yeaton & Sechrest, 1981). Variable implementation may diminish treatment effectiveness (e.g., Paine & Bellamy, 1982) and, in extreme cases, lead to treatment failures (e.g., Slavin, 1986; Stallings & Krasavage, 1986).

Important to successful large-scale dissemination of specific behavioral procedures has been the development of methods for assessing and analyzing the implementation of treatment agents, such as classroom teachers, so that treatment problems can be quickly and correctly diagnosed and appropriate corrective advice provided and implemented. A case in point has been the classwide peer tutoring procedure (Greenwood, Delquadri, & Carta, 1988) and its use in the public schools.

Classwide peer tutoring (CWPT) is a well-specified intervention in which tutor–tutee pairs work together on a classwide basis (Delquadri, Greenwood, Stretton, & Hall, 1983; Maheady & Harper, 1987). The program was developed over a series of single-subject and experimental-control group studies between 1981 and 1989 in which the program was validated, components refined, and replications extended to different subject areas, student populations, and ages (see reviews by Delquadri, Greenwood, Whorton, Carta, & Hall, 1986; Greenwood, Carta, & Hall, 1988; Greenwood,

This work was funded by grants from the Division of Innovation and Development, Office of Special Education Programs and Rehabilitation Services, U.S. Department of Education (G008730085) and the National Institute of Child Health and Human Development (HD01344) to the University of Kansas. This work reflects the opinions of the authors and not those of the funding agencies. This work is dedicated to the teachers, staff, and students of the Kansas City, Kansas, School District. The authors thank Judith Carta, Susan Fowler, Greg Harper, Debra Kamps, and Ilene Schwartz for their suggestions on early drafts of this manuscript. Thanks also are due to Janet Marquis for help with the statistical analysis. The authors also express their appreciation to the participating teachers at the Fairfax Elementary School in Kansas City, Kansas. Also deserving acknowledgement are Donald Moritz and Lowell Alexander, and the teaching staff of the Fairfax Elementary School, Kansas City, Kansas.

Send correspondence and reprint requests to Charles R. Greenwood, 1614 Washington Blvd., Kansas City, Kansas 66102.

Delquadri, & Hall, 1984; Greenwood, Maheady, & Carta, 1991). Briefly, when used to teach spelling, CWPT involves (a) weekly spelling word lists to be tutored, (b) new partners each week, (c) partner pairing strategies, (d) two teams competing for the highest point total, (e) tutee point earning contingent on correct responding, (f) tutors providing immediate error correction, (g) public posting of individual and team scores, and (h) social reward for the winning team (Greenwood, Delquadri, & Carta, 1988).

Teachers organize the academic content to be tutored into daily and weekly units and prepare materials for tutors and tutees to be used within the CWPT format. At the beginning of each week, all students in a class are paired for tutoring and these pairs are assigned to one of two competing teams. Tutoring occurs simultaneously for all tutor–tutee pairs involving all members of the class. This arrangement allows the classroom teacher to supervise and monitor students' responding (Greenwood, Carta, & Kamps, 1990). Tutees earn points for their team by responding correctly to the tasks presented to them by their tutors. Based on point totals, a winning team is determined daily and weekly. Tutor and tutee roles are highly structured to ensure that tutees receive rapid response trials in a consistent format and that tutors apply a standard error-correction procedure (e.g., Delquadri et al., 1983; Kohler & Greenwood, 1990).

For the small- and large-scale studies reporting the effectiveness of CWPT (Delquadri et al., 1986; Greenwood, Carta, & Hall, 1988; Maheady & Harper, 1987; Maheady, Sacca, & Harper, 1988), researchers reported that teachers and students sometimes varied standard implementation procedures. These variations included reduction in the number of CWPT sessions implemented per week (Dinwiddie, Terry, Wade, & Thibadeau, 1982; Greenwood, Delquadri, & Hall, 1989), reduction in the proportion of the class participating in CWPT sessions (Greenwood et al., 1989), addition, omission, and/or substitution of component procedures (Greenwood, Dinwiddie, et al., 1984; Greenwood, Maheady, & Carta, 1991; Kohler & Greenwood, 1990), provision of less-than-optimal material to be learned (Greenwood et al., 1987), and reduction in the fidelity of tutor–tutee interactions (Kohler & Greenwood, 1990; Kohler, Richardson, Mina, Dinwiddie, & Greenwood, 1985; Maheady & Harper, 1987). Because of these variations, the academic gains made by students in some CWPT programs may not have been optimal (e.g., Greenwood et al., 1989; Harper, Mallette, Maheady, & Clifton, 1990).

Based collectively on these observations and the continuing need for behavior analysts to conduct analyses of students' academic performance (Sulzer-Azaroff & Gillat, 1990), this investigation was designed to improve our assessment of CWPT implementation. To improve our consulting advice to teachers, we sought to improve our knowledge of natural implementation variations associated with less-than-optimal student outcome in CWPT. We also sought a basis for further experimental research on the implementation process, including why CWPT may break down and what components are more vulnerable. After assuring that CWPT was initially taught and implemented according to its published standards (Greenwood, Delquadri, & Carta, 1988), the following empirical questions were addressed:

To what extent were differences in students' CWPT outcomes, when defined by success and failure criteria, related to differences in implementation of CWPT? And, for success and failure student groups specifically, (a) what was the weekly opportunity to receive the CWPT program relative to program standards, number of sessions actually conducted, and number of sessions actually participated in by students (strength of treatment)?; (b) was the difficulty of the spelling words to be learned in CWPT each week set correctly according to program standards (fidelity of treatment)?; and (c) was the quality of tutor–tutee interactions within CWPT, as measured by point earning rates, in line with program standards (fidelity of treatment)?

METHOD

Participants, Setting, and Subject Matter

Five elementary school teachers (1, 2, 3, 4, and 5), 1 student teacher supervised by Teacher 3, and their students participated during the 1989–1990

school year. Teacher 5 was a long-time CWPT user and had participated in previous CWPT research studies (e.g., Delquadri et al., 1983; Greenwood et al., 1989). The other 4 teachers had heard about the program from other teachers at the school. These teachers, who volunteered to be trained in CWPT procedures, also agreed to allow researchers to visit the classroom and observe their implementation. They each received $100.00 for their participation.

All classes were located in the same inner-city school that served students from low socioeconomic levels. Because of the number of disadvantaged students, the school qualified for supplementary federal funding under the Education Consolidation and Improvement Act of 1981. These resources were used to provide a special compensatory educational program for any student with an academic delay defined by scores below the 49th percentile on a standardized achievement battery. As a result, students could attend a resource room where they were taught reading and mathematics by a special Chapter I teacher for as long as 2 hr per day.

The classes contained 19, 21 (Grade 2), 24, 23 (Grade 4), and 21 (Grade 5) students, respectively. The students attended the regular education program wherein the study took place during their regular spelling instruction. Teachers employed the Harcourt-Brace spelling curricula.

Design

A nonexperimental, multimeasure, clinical replication design was used (Barlow, Hayes, & Nelson, 1984; Greenwood et al., 1987). A clinical replication is differentiated from a direct or systematic replication in its focus on the replication of a well-defined treatment procedure with a large number of participants for the purpose of establishing the generalizability of its effectiveness in an applied setting (Barlow et al., 1984). The design is appropriately applied following a series of earlier studies focused on technique building as a means of exploring instances of nonimprovement in relationship to variations in treatment implementation and/or subject characteristics. The design is used to generate hypotheses for future experimental study concerning specific implementation conditions associated with subjects' failure to achieve the expected treatment outcome (Barlow et al., 1984). This goal of hypothesis generation is somewhat similar to the usual testing phase in a functional analysis design, whose purpose is to explore the effects of alternative environmental conditions on an aberrant behavior prior to conducting the actual experimental analysis (cf. O'Neill, Horner, Albin, Storey, & Sprague, 1990).

In the present study, we defined students' weekly spelling achievement in the tutoring program in terms of four conditions that reflected key program standards. These standards reflected the appropriateness of the material to be learned prior to tutoring and mastery of the material after tutoring. Thus, students were counted members of the *success group* (SUC) during a week in which they both (a) were challenged by the material to be learned (pretest less than 40%) and (b) had mastered 80% or more of the material at week's end (Greenwood, Delquadri, & Carta, 1988). These particular criteria were based on prior research (e.g., Greenwood, Delquadri, & Hall, 1984). They ensured the spelling words chosen for the week were appropriately difficult for most students, and they provided evidence of an adequate implementation of CWPT.

Three remaining groups (challenged/undermastery, underchallenged/mastery, and underchallenged/undermastery) also were defined in terms of direct relationships to implementation factors and program standards. Each of these groups, however, represented some deviation from program standards. For example, students qualified for the challenged/undermastery group (C/UM) during a week in which they were provided spelling words of appropriate difficulty but failed to achieve at least 80% at posttest. Prior experience indicated that these students had very likely encountered low-strength and/or low-quality tutoring (e.g., Greenwood, Dinwiddie, et al., 1984; Harper et al., 1990).

Students qualified for the underchallenged/mastery group (UC/M) in a week in which they were provided with spelling words that were too easy, as measured by a pretest score exceeding 40% correct, and who scored at least 80% by the posttest. This outcome meant that students' posttest achieve-

ment gain had been reduced due to floor and ceiling effects (Greenwood et al., 1987).

The underchallenged/undermastry group (UC/UM) represented the combination of spelling items that were too easy (i.e., greater than the 40% pretest standard) and failure to reach the 80% posttest standard. In this case, students' weekly outcomes may have been reduced by ceiling effects and/or reductions in the strength or lowered fidelity of CWPT.

Measurement Model

Multiple measures of students' performance during CWPT and of teachers' program implementation were employed to reflect student achievement as well as strength and fidelity of treatment. Student achievement measures included weekly spelling pre- and posttests. Strength of treatment was assessed in terms of (a) the weekly opportunity to receive CWPT and (b) each student's actual presence and participation in CWPT sessions. Fidelity of treatment was assessed in terms of (a) a CWPT procedural checklist, (b) points earned by students during daily sessions, and (c) tutor–tutee procedural calibration probes.

Student achievement. Students' spelling achievement was assessed using 20-item pre- and posttests reflecting the material to be learned in a week as in prior studies (Greenwood, Dinwiddie, et al., 1984; Greenwood et al., 1987). Teachers dictated the words to be spelled, one at a time, and students attempted to write them. The tests were corrected by the teachers and scored in terms of percentage correct. The reliability of these tests has been high in prior studies. For example, Pearson *r*s of 0.88 and 0.97 between teacher and consultant scoring of the same tests with no significant differences between spelling test mean scores were reported for 2 separate years (Greenwood et al., 1987).

Strength of CWPT treatment. The opportunity for students to receive CWPT and students' participation in these opportunities were monitored. The number of daily sessions implemented per week and for which data (i.e., points earned) were available for at least 1 student defined the occurrence of a CWPT opportunity. The possible number of CWPT opportunities ranged from zero to four per week. A CWPT opportunity that was actually attended and participated in by a specific student was defined as CWPT participation. It reflected CWPT sessions missed by individual students due to absences or assignment to other instruction (e.g., Chapter I). Participation was indicated by an individual student's tutoring data (e.g., points earned) on a specific day when a session had been held. The probability of CWPT participation was computed by dividing the weekly rate of CWPT participation by the weekly rate of CWPT opportunities. Because the evidence for CWPT opportunities and participation was a permanent product record on students' point charts, reliability was assumed to be 100%.

Fidelity of CWPT treatment. The three-category CWPT procedural checklist developed in prior research (Greenwood et al., 1987, 1989) was used to certify each teacher as trained (Week 4) and to assess maintenance of the quality of their classroom implementation of the CWPT program (Week 18). In prior research, teachers achieved fidelity percentage means of 82.7% (Year 1) and 90.6% (Year 2) with a range of 57% to 97% over all checks (Greenwood et al., 1987).

The checklist was administered to each teacher by project staff members unannounced on randomly selected days. Items on the checklist were scored as either present or absent in terms of three implementation categories. These were (a) presence of CWPT materials (7 items), (b) teacher use of CWPT procedures in correct sequential order (15 items), and (c) the tutoring interactions of a randomly selected pair, also in sequential order (14 items). Adequate implementation was defined by a composite score of 85% or higher on this checklist.

The number of points each student earned during a tutoring session and reported to the teacher during daily sessions was used as an index of tutoring fidelity. Reliabilities on point earning and point reporting have ranged from 88.0% to 98.0% based on percentage agreement statistics (Maheady & Harper, 1987). Prior research indicated that student point earning during individual CWPT sessions was a valid indicator of spelling practice (i.e.,

the number of word trials completed and corrected). It is also a global indicator of the number of different spelling words practiced during tutoring sessions (e.g., Greenwood, Dinwiddie, et al., 1984; Kohler & Greenwood, 1990; Maheady & Harper, 1987). Higher point totals reflected completion of more word trials and practice distributed across more spelling words.

Tutor–tutee procedural calibration probes were made of the core tutor–tutee behaviors (e.g., Kohler & Greenwood, 1990) when point earning data (e.g., outliers) suggested a problem. The probes were designed to assess and diagnose the problem in terms of a specific breakdown in the core tutor–tutee behaviors. Observed in real time were the number of spelling words actually presented to and attempted by tutees, the words correctly written, the words written in error and accurately corrected by the tutor, and the number of points actually earned and reported to the teacher after the session.

These probes were conducted during 2 weeks with 6 students in Class 5. Probes were necessary because of impossibly high point totals for 3 students and the failure of 3 other students to increase their low point totals over days in the week. In prior research, reliability on core CWPT tutoring behavior probes averaged 97%, ranging from 64% to 97% (Kohler & Greenwood, 1990).

The spelling pre- and posttests, CWPT opportunities, CWPT participation, and point earning measures were collected by the classroom teachers using the standard data collection procedures described in the CWPT manual (Greenwood, Delquadri, & Carta, 1988). Each day these data were compiled into a classroom data base using a laptop computer data entry program developed specifically for this purpose (i.e., the CWPT Support Program; Greenwood, Finney, Terry, & Arreaga-Mayer, 1990). Each week's CWPT implementation data were uploaded to an IBM-PC® compatible desk-top computer by the investigators and accumulated for statistical analysis. The procedural checklist and calibration probes were conducted by project staff members and entered in another data base.

Available for analysis after 19 weeks in each class were 288, 318, 306, 305, and 253 weekly records for each student for Teachers 1, 2, 3, 4, and 5, respectively, or a total of 1,462 records for 110 students, when reduced by missing data (i.e., 110 students × 19 weeks is equal to 2,090 records compared to 1,462. The difference of 628 records represents the effects of incomplete data.).

Reliability

The traditional indices of agreement and reliability were not assessed in the current study for reasons related to the purposes of the research. Because our goal was to observe CWPT program variation over time under natural conditions, reliability checks were not used in an effort to reduce any unanticipated positive effects of such checks on teachers' maintenance of program quality (e.g., Hartmann & Wood, 1982). Furthermore, formal reliability checks in the context of the program's usual classroom operation do not occur and represent an additional research requirement.

However, to ensure continuity with prior research, all measures in the present study were calibrated against class means and ranges for the same measures in past research. None exceeded these expected parameters. All measures employed in the study had been previously validated and found reliable in earlier experimental research studies (e.g., Greenwood, Dinwiddie, et al., 1984; Greenwood et al., 1987; Kohler & Greenwood, 1990; Maheady & Harper, 1987).

Procedures

A field trial of CWPT was conducted in a single school. The aim was to monitor teachers' variations in implementation under natural conditions as the program became part of the total set of practices and procedures employed at the school. After training teachers and students to implement the program according to usual standards, minimal efforts were made to further influence, improve, or shape the directions of the program over time.

CWPT teacher training. Project staff members trained the teachers to implement the CWPT procedures during December. Implementation occurred immediately thereafter through May. Teach-

ers initially read a CWPT program manual that described the procedures (Greenwood, Delquadri, & Carta, 1988) and then discussed with their consultant-trainer the necessary changes to be made in current classroom practices (e.g., Maheady, Harper, Mallette, & Winstanley, 1991). They also learned to use the CWPT support program for data entry purposes.

The program standard requires CWPT sessions to be conducted four times per week for two 10-min tutoring sessions per day, Monday through Thursday; all teachers agreed to implement this schedule. Friday is used for pre- and posttesting. After the necessary planning and preparation of materials, staff members assisted the teachers in initiating the program in their classrooms.

The spelling words used in the program were derived from the school's scope and sequence goals as well as the grade-level spelling and reading curriculum. Teachers prepared 20-word lists, one for each week, in a sequence compatible with their plans for teaching it.

Throughout the study, project staff members picked up weekly data and responded to any problems expressed by teachers. Project staff members also held four monthly meetings that all teachers attended after school in order to review their progress, clarify any procedures, and discuss problems.

Classroom implementation. Teachers trained their students to implement CWPT in four short lessons in which procedures were described, modeled, role-played, and then practiced in isolation as directed in the manual (Greenwood, Delquadri, & Carta, 1988). These lessons covered (a) the CWPT game, (b) winning and losing teams, (c) working with a partner, and (d) being a peer tutor, all prior to the first full CWPT session.

CWPT sessions. At the beginning of each week, the pairing of tutoring partners for the week was completed randomly by the teacher. At the beginning of each daily tutoring session, the teacher reminded the students to check the partner chart posted in the classroom for their partner assignment. This chart displayed the partners for the week, their team membership, and which partner served as tutor first. The teacher then instructed the students to move to their partners as she cued the NEC-8300 lap-top computer to begin timing the first 10-min tutoring period.

Each tutor presented the first word from the list of words to be learned by the tutee. The tutee then responded by writing and saying the word. The tutor then checked the response by comparing it to the correct answer on the list. When an error occurred, the tutor immediately provided the correct answer and then required the tutee to practice it by writing it three times. Tutees earned two points for each correct answer and one point for correcting an error.

At the end of the first 10-min period, the tutor and tutee traded roles and a second 10-min period was completed. Following the second period, a 5-min period was used by students to report orally the total points each had earned; these were posted on their team's chart. Individual points were summed and team totals announced. The winning team was applauded and the losing team was encouraged to work harder in the next session. The teacher then moved on to the next activity.

Quality of implementation. Staff used the procedural checklist to certify teachers as trained and to assess maintenance of their implementation. All teachers' implementations of CWPT were rated at or above the 85% minimum criterion on the CWPT procedural checklist during Week 4. These percentages ranged from 85% (Student Teacher 3) to 100% (Teacher 4). These assessments also indicated that teachers were devoting the correct total time to CWPT (M = 27.4 min; range, 25 to 31). At Week 18, 3 of the 5 teachers maintained levels of implementation fidelity above the 85% minimum with the exception of Teacher 1, at 81%, and Teacher 3, who did not implement the program that week. The time devoted to CWPT sessions continued to be adequate (M = 32.5 min; range, 25 to 45).

Data Analysis

To address the research questions, a combination of analytic strategies was employed. First, a description of the frequency with which students experienced the four weekly outcomes was computed

across the 19 weeks. This distribution was as follows: 47.9%, 29.5%, 19.3%, and 3.3% for the UC/M group, SUC group, C/UM group, and UC/UM group, respectively. Seventy-seven students (70%) had experienced two outcome groups, 52 (47%) had experienced three groups, and only 20 (18%) had experienced all four groups. Thus, it was more common for students to fall into the UC/M group, whereas they were least likely to fall into the UC/UM group, and only a relative few had experienced all four groups across their 19-week participation in the program.

Second, a statistical analysis was performed over all teachers and students in order to reveal any systematic implementation differences associated with the four outcome groups. Several considerations shaped the statistical analysis, including the unbalanced distribution of students to outcome groups.

Because of the likelihood of serial correlation within students' data repeatedly measured across weeks and violation of the independence assumption in parametric statistical analysis (e.g., Busk & Marascuilo, 1988; Jones, Vaught, & Weinrott, 1978), steps were taken to limit this problem. First, the number of repeated measures was reduced by collapsing the data by weeks and using students rather than weeks as the unit of analysis. Second, multiple comparisons between pairs of outcome group means were made using *t* tests for dependent data. These tests accommodated the fact that repeated measurements made on subjects were correlated (Dixon, Brown, Engelman, & Jennrich, 1990). Testing pairs of means allowed a sensitive evaluation of differences, because each test was based on all the data available for each student, whereas alternative methods of simultaneously comparing all four means (e.g., ANOVA for repeated measures) would have drastically reduced the number of students in the analysis to only those with complete data. An alpha level of 0.05 was used for all paired comparisons.

Finally, a nonstatistical analysis of each teacher's CWPT implementation was conducted using simple graphic displays to represent the relationships between strength of treatment, fidelity of treatment, and student achievement in each individual classroom.

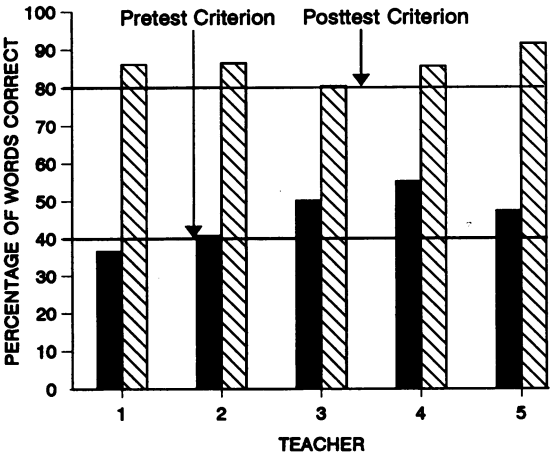

Figure 1. Spelling accuracy pretest and posttest means by teachers/classrooms combined over all program weeks.

RESULTS

The average spelling pretest (M = 46.0%) to posttest (M = 85.8%) mean gain combined over all classes and weeks was 39.7%. As illustrated in Figure 1, the individual class gains ranged from a low of 30.3% (Classes 3 and 4) to a high of 49.5% (Class 1). These improvements in spelling accuracy compared favorably to those previously reported (e.g., Greenwood et al., 1987) and confirmed the extent to which these effects replicated those reported in prior investigations.

Because of the criteria used to classify students into the four outcome groups, the groups differed significantly at pretest, posttest, and in terms of gain, with the exception of the posttest comparison between the C/UM group versus the UC/UM group (see Table 1). The largest spelling test gains were made by the SUC group (66.2%), followed by the C/UM group at 38.2%, the UC/M group at 28.2%, and the UC/UM group at 0.9%.

Implementation Factor Differences by Outcome Groups

The success group. The SUC group worked on material that was challenging (pretest M = 25.9%)

Table 1
Spelling Accuracy Means by Outcome Groups

Measures	Outcome groups			
	SUC	UC/M	C/UM	UC/UM
Number of students	88	97	66	31
Pretest	25.9	67.2	19.1	61.3
	(8.2)	(11.7)	(8.5)	(13.8)
Posttest	91.8	95.4	59.5*	61.7*
	(6.0)	(4.0)	(10.7)	(13.6)
Gain	66.2	28.2	38.2	0.9
	(8.7)	(10.2)	(17.2)	(20.0)

Note. Means with asterisks are not significantly different from each other. Standard deviations are given in parentheses. SUC = success (pretest < 40%, posttest ≥ 80%); C/UM = challenged/undermastery (pretest < 40%, posttest < 80%); UC/M = underchallenged/mastery (pretest > 40%, posttest ≥ 80%); UC/UM = underchallenged/undermastery (pretest > 40%, posttest < 80%).

and they achieved mastery (posttest M = 91.8%). They also received the best overall implementation of CWPT in terms of strength and fidelity of treatment (see Table 2). They had high weekly opportunities to receive CWPT (M = 3.0 days/week), the highest probability of participating in these opportunities (p = .92), and an ascending trend in daily point earnings, with means ranging from 63.6 points on Monday to 122.2 on Thursday.

The challenged/undermastery group. The C/UM students were challenged by the material (pretest M = 19.1%), but they did not reach mastery by week's end (posttest M = 59.5%). Although they received an equal number of CWPT sessions per week (M = 3.0) compared to the SUC group, the distinguishing feature of their program was a significantly lower probability of participation in sessions held (p = .83) combined with lower rates of points earned during tutoring sessions. The C/UM group, as did the SUC group, had an ascending point earning trend over the week; however, it was systematically lower, ranging from 42.6 on Monday to 80.9 on Thursday.

The underchallenged/mastery group. The UC/M students worked on material that was too easy for them (pretest M = 67.2%) according to CWPT criterion, and after a week, they did reach mastery (posttest M = 95.4%). Although this group experienced statistically fewer CWPT sessions per week (M = 2.8) than did the prior two groups, their participation in sessions was high (p = .91).

Table 2
Implementation Factor Means by Outcome Groups

Variables		Outcome groups			
		SUC	UC/M	C/UM	UC/UM
Implemented sessions	*M*	3.0	2.8	3.0	2.1
	SD	1.0	0.9	0.8	1.3
	N	88	97	66	31
Probability of participation	*M*	.92	.91	.83	.63
	SD	.17	.14	.22	.27
	N	86	95	66	25
Points—Monday	*M*	63.6	99.2	42.6	90.0
	SD	34.7	51.6	25.6	89.9
	N	86	95	66	24
Points—Tuesday	*M*	105.7	137.5	67.4	98.7
	SD	65.4	62.5	45.7	87.6
	N	79	93	59	18
Points—Wednesday	*M*	117.7	160.1	69.5	102.9
	SD	66.1	73.5	42.4	72.9
	N	68	81	46	14
Points—Thursday	*M*	122.2	160.1	80.9	—
	SD	64.4	89.3	53.4	—
	N	60	69	38	2

Note. Abbreviations as in Table 1.

They also earned a significantly higher number of points on Monday and throughout the week, means ranging from 99.2 to 160.1. Unlike the first two groups, whose points peaked on Thursday, their point earning trend peaked on Wednesday. This was the only group to show this trend.

The underchallenged/undermastery group. The UC/UM students worked on material that was too easy (pretest M = 61.3%) and did not attain mastery (posttest M = 61.7%). Only 2 of 15 comparisons involving this group were significantly different from those of the other groups (Table 2). However, this group experienced the combination of low strength and low program fidelity, which included the lowest rate of CWPT implementation (M = 2.1 sessions), the lowest probability of CWPT participation (p = .63), and an incomplete, relatively flat point earning trend, with means ranging from 90.0 to 102.9 over the week (see Table 2).

Implementation Factor Differences by Teachers and Classrooms

The complex relationships between students' outcome and the strength and fidelity of treatment overall were even clearer when examined by classroom. The students of Teachers 3 and 4 made the least spelling gains, whereas the students of Teacher 1 made the most (see Figure 1). The implementation data suggested several reasons for this finding.

Teacher 4's implementation was uniquely characterized by the highest proportion of UC/M students each week (M = 61%) combined with a relatively high-strength (CWPT opportunities ranged from 2.3 to 3.4 across groups, and the probability of student participation ranged from .73 to .95 across groups) and a high-fidelity (daily point earning) program (see Figure 2). Replicating the overall statistical analysis, strength of treatment was lowest for the UC/UM group in this class compared to the other outcome groups. Also replicated was the relationship between the high number of UC/M students and this group's relatively high rates of daily point earning.

Like Teacher 4, Teacher 3's program indicated a sizable proportion of UC/M group students (M = 49%), and she also had the largest C/UM group (M = 24%) in combination with the lowest strength of treatment. Of the 5 teachers, Teacher 3 implemented CWPT least often, ranging from 1.4 to 1.8 sessions across groups, and her students were most likely to be absent from these sessions, with probabilities ranging from .52 to .83 across groups (see Figure 3). Point earning data for her students were often incomplete after Tuesday of the week, further reflecting the reduced levels of CWPT implementation and lowered student participation. Point earning trends in each group were relatively flat or declining, suggesting low tutoring fidelity and reduced spelling word practice.

In contrast, Teacher 1's students made the largest gains in the program. Teacher 1 had the highest proportion of her class falling into the SUC group each week (M = 45%), and none of her students ever fell into the UC/UM group. Teacher 1 also implemented the highest strength program, with most weekly session means ranging between 3.6 and 3.7 over groups; the probability of students participating in these sessions was also very high, ranging from .92 to .97 across groups (see Figure 4). Point earning data also indicated a high-fidelity program. Point earning trends were accelerating for all groups over the week, with relatively small differences between the four groups in their point earnings (see Figure 4).

Tutor–tutee interaction. Procedural calibration observations of selected students confirmed a number of additional facts about the relationship between individual students' point earnings and the fidelity of their tutoring interactions. For example, Student 13 in Class 5 was targeted for procedural calibration assessments because she was consistently the class outlier with respect to point totals. The check revealed that she obtained 100% on the pretest and on a Monday completed 77 words with only two errors. Her actual point total was 154. However, she reported 254. On Tuesday, she wrote 81 words with no errors and earned a total of 162 points, but she reported 262. Similarly, inflated point reports were made by Students 6 and 16 in this same class.

Students in Class 5 who performed lowest at pretest also tended to have a number of unique problems related to the quality of their tutoring

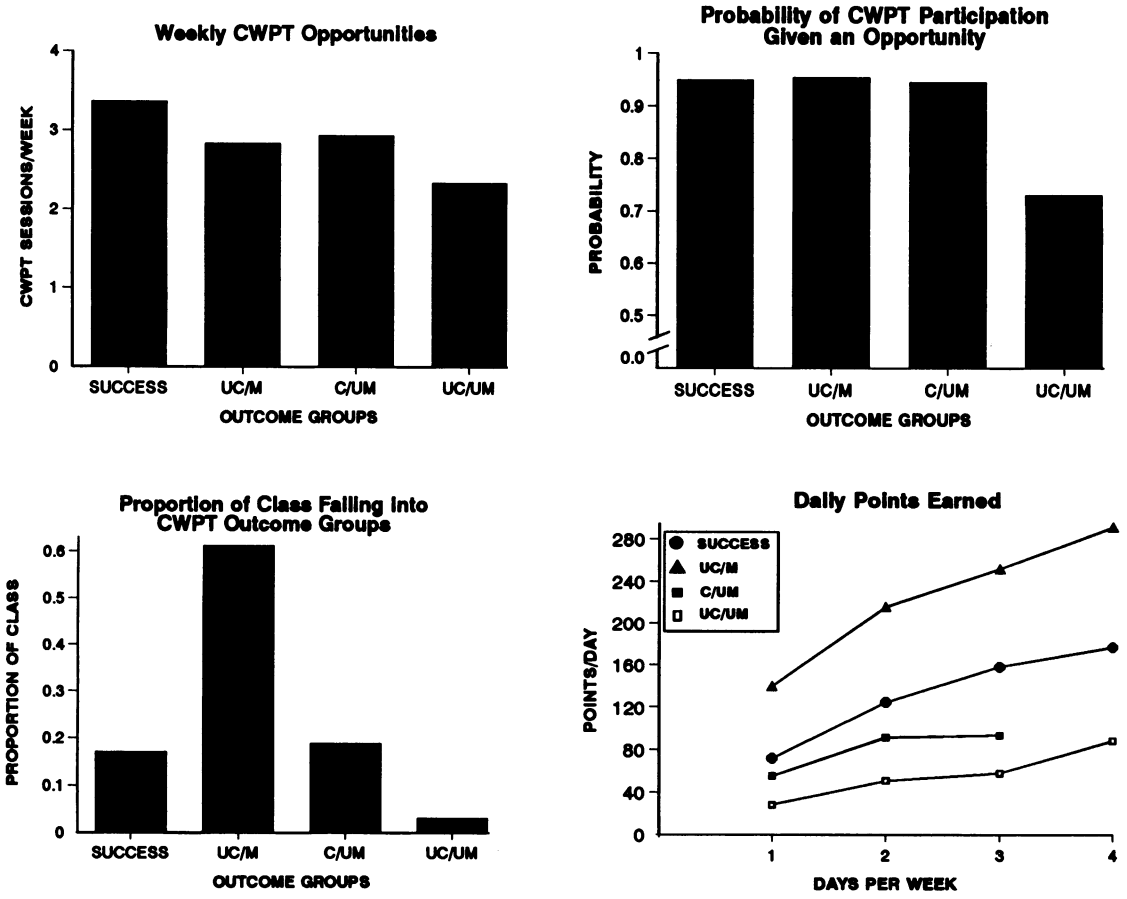

Figure 2. Implementation data summary for Teacher 4. (Abbreviations are C/UM = challenged/undermastery, UC/M = underchallenged/mastery, and UC/UM = underchallenged/undermastery where success = pretest < 40%, posttest ≥ 80%; C/UM = pretest < 40%, posttest < 80%; UC/M = pretest > 40%, posttest ≥ 80%; and UC/UM = pretest > 40%; posttest < 80%.)

interactions (i.e., a low number of words correct and covered during the session, and thus, low point earning). For example, Student 21 wrote 47 words on a Monday, of which 31 were in error. Observation indicated that none of his errors were corrected by the tutor. On Tuesday, he wrote 25 words with seven errors. In this instance, the same tutor applied the error-correction procedure to only two of these seven errors. Similar failures of tutors to apply the error-correction procedures were also observed for Students 11 and 22. Also noted for Student 11 was the fact that his tutor had a very slow presentation rate, and the pair did not engage in the tutoring task for the entire 10 min.

DISCUSSION

Consistent with the clinical replication design, we examined variations in students' spelling outcomes and their relationships with variations in strength and fidelity of implementation. After a series of experimental studies that developed components and that validated the efficacy of CWPT, our purpose was to improve the ability to diagnose CWPT program implementations as a basis for improving teachers' implementation, students' academic performance, and the quality of implementation advice.

Results indicated that students in each class made

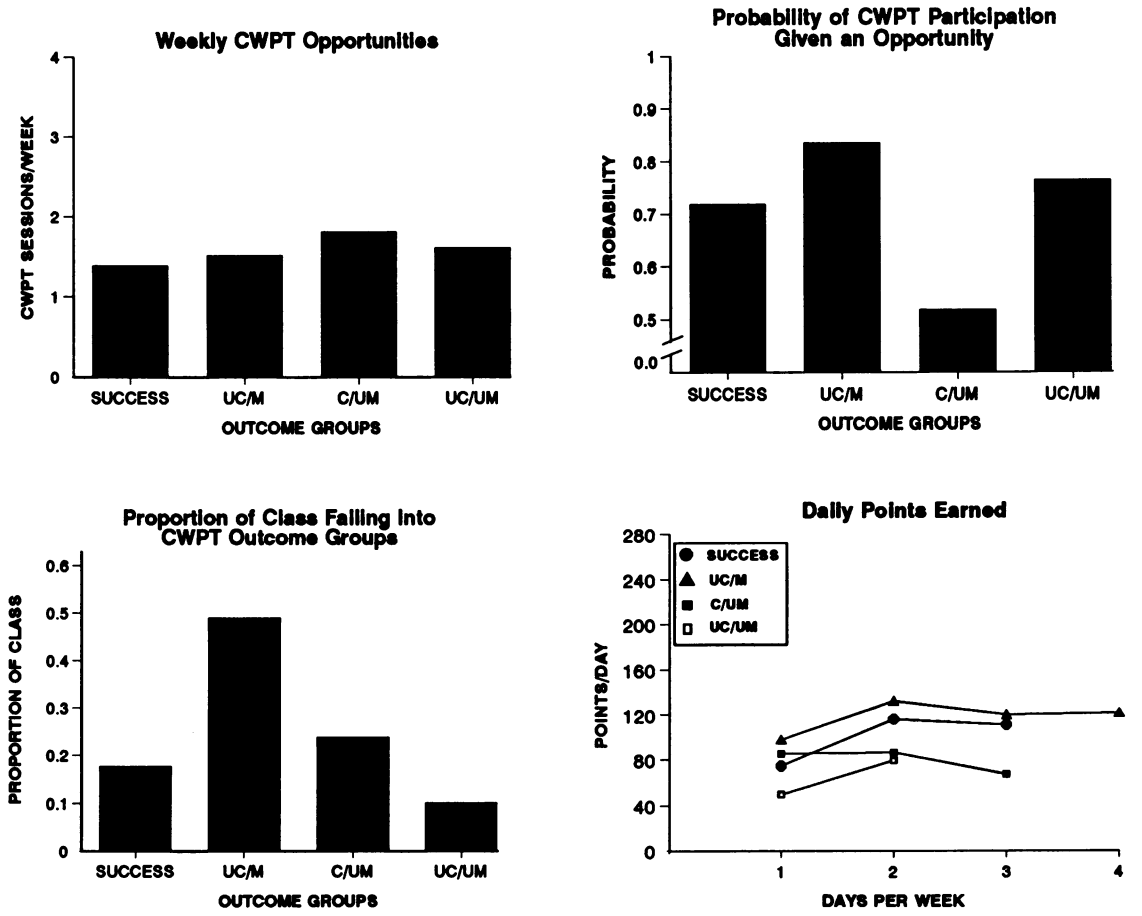

Figure 3. Implementation data summary for Teacher 3. (Abbreviations as in Figure 2.)

educationally important gains in spelling accuracy, and CWPT again was demonstrated to be a robust procedure (Greenwood et al., 1987). Also, as reported in prior research (e.g., Greenwood et al., 1989), neither the teachers' implementation of CWPT nor students' spelling gains were considered optimal relative to program standards. The statistical analysis as well as analyses by teacher revealed that multiple implementation differences were associated with spelling outcomes.

We found that variations in students' spelling outcomes were associated with (a) reductions in strength of treatment (i.e., CWPT opportunity and student participation) and (b) low program fidelity (i.e., unchallenging spelling words at pretest), as well as (c) low point earnings during tutoring. A synthesis of the problems identified from these analyses, their implications, and suggested procedures for correcting them are summarized in Table 3. This synthesis represents a basis for future experimental research on CWPT program evaluation, diagnosis, and advice.

Teacher 4's implementation illustrated how a single fidelity factor (e.g., material too easy) could affect minimal learning within an otherwise high-strength, high-fidelity CWPT program. Teacher 3's implementation represented the interaction of both low-strength and low-fidelity problems, and her history in the program was most instructive. Although this teacher volunteered to participate, her student teacher actually initiated CWPT in the classroom; the student teacher's implementation was checked on Week 4. The program was then taken over by Teacher 3 several weeks later; the accuracy

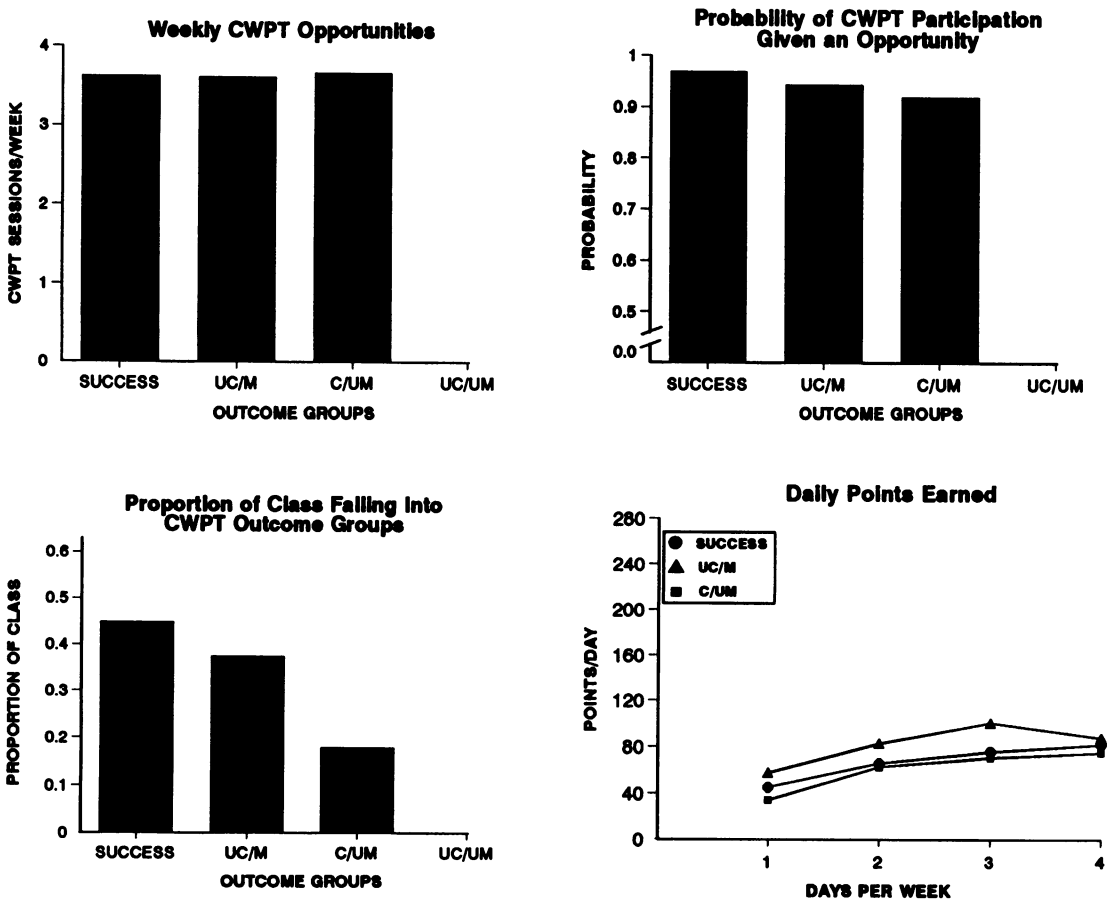

Figure 4. Implementation data summary for Teacher 1. (Abbreviations as in Figure 2.)

of her implementation could not be assessed because she did not implement the program on Week 18, when the second fidelity check was conducted. Lack of continuity in treatment agents and program implementation is not an uncommon problem in schools and other organizations.

This class also contained the highest number of students sent to Chapter I reading and mathematics sessions held outside of the classroom; in addition, some students received speech therapy. Up to 14 of her 22 students were gone for some portion of each day. This created a scheduling problem for CWPT that was never addressed adequately by the teacher. She used the program with whoever was in the classroom on those days when she decided to use the program. Thus, her problem was a combination of conflicting instructional demands on students' time and noncompliance with the scheduling aspects of the standard CWPT program. These factors, combined with the high proportion of students given unchallenging spelling content each week, further reduced the viability of this program and, acting together, appeared to limit what students were able to learn.

Also instructive was the performance of Teacher 1, whose class maintained the highest levels of spelling gain and who also provided the strongest treatment in terms of CWPT opportunity and student participation. Teacher 1 was able to schedule CWPT so as not to compete with Chapter I. She also used her pretest information much more effectively as a basis for selection of the material to be learned each week. Even so, her procedural checklist data at Week 19 indicated that some components of the program had been dropped and that her score was just below the minimum 85%

Table 3
Synthesis of Implementation Variations, Implications, and Corrective Actions

Implementation problem	Teacher(s)	Implication	Corrective Action
Underchallenged students (pretest ≤ 40%)	2, 3, 4, 5	Less than optimal weekly spelling gains	Increase the number of challenging words each week
Words are too difficult at pretest (pretest > 20%)	C/UM group	Less likely to achieve mastery	Monitor tutoring interactions closely; reduce number of words to increase practice per word
Teacher implements fewer CWPT sessions than available (less than four per week)	3, and UC/UM group	Reduction in the opportunity to learn words	Check manual and renew program commitment
The discrepancy between sessions implemented and participated in by students is too large	3, and UC/UM group	Reduction in students' CWPT participation	Assess reasons for absences. If schedules conflict, review goals. If school absences, contact parents
Tutoring point earning is too low	C/UM group	Reduced word trials; reduction in word coverage	Assess tutoring interactions, check for delays and error correction
Tutoring point earning is too high	UC/M group	Underchallenged students, inaccurate point reporting	Supply challenging word lists, check reliability of point reporting, check error correction
Tutor does not correct tutee's errors	UC/UM group	Tutor training was insufficient; teacher is not using bonus points to maintain error correction	Retrain specific pairs; review bonus point procedures and rationale, use bonus points contingent on error correction
Tutor's word presentation rate is too low	C/UM group	Tutor is failing to monitor tutee; tutee's practice and content coverage is reduced	Retrain tutor; use bonus points contingent on increased word presentations and tutee responses

level. Teacher 1 no longer praised students or used bonus points, and did not praise the winning or encourage the losing team. She also experienced some behavior problems after tutoring during the point reporting period. Also, her sessions had increased in length by at least 15 min beyond the 30-min session standard.

A side effect of most students in a classroom working on unchallenging material (e.g., Teachers 4 and 5) appeared to be inflation of the overall classroom point economy, presumably because more students were practicing words they already knew. Anecdotally, in Class 5 it seemed that an inflated point economy and reduction in teacher monitoring of individuals may have led to instances of point cheating, tutors' failing to identify and implement error correction when errors occurred, and a reduction in teacher monitoring of these events and provision of consequences to prevent them from happening. This outcome deserves further research.

The reduction in CWPT sessions and student participation (strength of treatment) observed in this study was a larger threat to student outcome than we had previously anticipated, particularly for the two undermastering groups (C/UM and UC/UM), and also considering that these were volunteer teachers who were interested and motivated to use the program. Reductions in weekly tutoring sessions in most cases may have reflected the fact that because most students in each class were unchallenged by the weekly material, they only needed three rather than the standard four tutoring sessions per

week to master their spelling words. In the worst case, reductions in CWPT weekly opportunities may have reflected loss in teacher commitment to CWPT or an inability to plan implementation of the classroom program given so many competing demands on students' time (e.g., to attend the Chapter I program).

Several limitations were imposed by the design used in this investigation. Because an experimental design was not used, the current results are descriptive and correlation in nature. Thus, these findings await validation in future experimental studies of CWPT implementation. The unbalanced data set created a small sample size, combined with large standard deviations within the statistical comparisons involving the UC/UM group, and consequently most were not significant. However, the UC/UM group means reflected the lowest strength and lowest fidelity estimates of implementation over all students, and these values and their differential association with student outcomes were cross validated in each of the five individual classroom replications. Thus, the reliability of these relationships was supported.

To reduce the effects of the research context on teachers' implementation, we did not collect the usual reliability information on our dependent variables. Instead, we checked the observed outcomes against prior parameters (e.g., means and standard deviations) in our earlier published studies, and we found a high degree of replication. In specific cases in which individual students' scores exceeded expected limits on measures (i.e., Class 5), we conducted procedural calibration probes as a means of diagnosing and validating the specific tutoring conditions associated with these unreasonable values.

The implications of these findings for those with administrative responsibility for such programs is that monitoring of strength as well as fidelity of treatment should be included within quality-control assessment plans. We also need to conduct research designed to further our understanding of the factors that affect teachers' utilization, and in turn, how these factors affect efficacy.

The future success of programs such as CWPT may also depend heavily on social validity factors as well as on quality-control procedures (e.g., monitoring, feedback, and problem-solving advice). Concerning social validity, there appears to be an increasing link between implementation issues and the social invalidity of behavioral programs that deserves consideration (e.g., Schwartz & Baer, 1991). These authors point out that consumers may "not implement . . . some or all of the program's procedures . . . despite generally positive responses . . ." (p. 190) or intent to do so, because of low acceptability. A question then is, to what extent is variable implementation attributable to low acceptability, on the one hand, versus more traditional factors such as the quality of initial training and quality control procedures, on the other? This interesting relationship remains an area for future research.

Concerning quality control, an emerging alternative to traditional methods (such as manuals, materials, training to criteria, feedback, and human consultants) is computerized systems that support assessment of weekly outcomes, diagnosis of problems, and corrective actions known to be commensurate with the problem (Terry, Greenwood, Arreaga-Mayer, Walker, & Finney, 1990). We are currently investigating the effects of a CWPT expert system—a computerized teacher consultant that provides weekly program assessment, diagnosis, and implementation advice to teachers. The computer program uses a diagnostic strategy based on the data contained in this and prior CWPT studies (Greenwood, Terry, & Arreaga-Mayer, 1991). A preliminary AB design indicated that the computer program was effective in identifying unchallenging material as a problem when introduced into an ongoing CWPT program and that its recommendation to increase spelling word difficulty reduced the frequency of this problem and increased students' weekly academic gains (Greenwood, Terry, Arreaga-Mayer, & Finney, 1991).

In addition, the correction of faulty program implementation and research on factors affecting program use in applied settings may depend on the development of assessment methods and measurement models more sensitive and complex than only those used in the initial training and certification

of its implementers. Assessments that combine data on strength and fidelity of treatment with client performance and client outcome may be needed. The data and analyses in this study provide just such a basis for future experimental research on the use of programs like CWPT and for improving the procedures by which implementation advice is made available for effective use by teachers.

REFERENCES

Barlow, D. H., Hayes, S. C., & Nelson, R. O. (1984). Clinical replication. In D. H. Barlow, R. O. Nelson, & S. C. Hayes (Eds.), *The scientist practitioner: Research and accountability in clinical and educational settings* (pp. 290–346). New York: Pergamon.

Busk, P. L., & Marascuilo, L. A. (1988). Autocorrelation in single-subject research: A counterargument to the myth of no autocorrelation. *Behavioral Assessment,* **10,** 229–242.

Carta, J. J., & Greenwood, C. R. (1989). Establishing the integrity of the independent variable in early intervention programs. *Early Education and Development,* **1,** 127–140.

Delquadri, J., Greenwood, C. R., Stretton, K., & Hall, R. V. (1983). The peer tutoring game: A classroom procedure for increasing opportunity to respond and spelling performance. *Education and Treatment of Children,* **6,** 225–239.

Delquadri, J., Greenwood, C. R., Whorton, D., Carta, J., & Hall, R. V. (1986). Classwide peer tutoring. *Exceptional Children,* **52,** 535–542.

Dinwiddie, G., Terry, B., Wade, L., & Thibadeau, S. (1982). *The effects of peer tutoring and teacher instructional allocation on academic achievement outcomes.* Poster presented at the eighth annual convention of the Association for Behavior Analysis, Milwaukee, WI.

Dixon, W. J., Brown, M. B., Engelman, L., & Jennrich, R. I. (Eds.). (1990). *BMDP statistical software manual* (Vol. 1). Berkeley, CA: University of California Press.

Greenwood, C. R., Carta, J. J., & Hall, R. V. (1988). The use of classwide peer tutoring strategies in classroom management and instruction. *School Psychology Review,* **17,** 258–275.

Greenwood, C. R., Carta, J. J., & Kamps, D. (1990). Teacher versus peer-mediated instruction. In H. Foot, M. Morgan, & R. Shute (Eds.), *Children helping children* (pp. 177–206). Chichester, England: Wiley.

Greenwood, C. R., Delquadri, J., & Carta, J. J. (1988). *Classwide peer tutoring (CWPT).* Delray Beach, FL: Education Achievement Systems.

Greenwood, C. R., Delquadri, J., & Hall, R. V. (1984). Opportunity to respond and student academic performance. In W. Heward, T. Heron, D. Hill, & J. Trap-Porter (Eds.), *Behavior analysis in education* (pp. 58–88). Columbus, OH: Charles E. Merrill.

Greenwood, C. R., Delquadri, J., & Hall, R. V. (1989). Longitudinal effects of classwide peer tutoring. *Journal of Educational Psychology,* **81,** 371–383.

Greenwood, C. R., Dinwiddie, G., Bailey, V., Carta, J. J., Dorsey, D., Kohler, F., Nelson, C., Rotholz, D., & Schulte, D. (1987). Field replication of classwide peer tutoring. *Journal of Applied Behavior Analysis,* **20,** 151–160.

Greenwood, C. R., Dinwiddie, G., Terry, B., Wade, L., Stanley, S., Thibadeau, S., & Delquadri, J. (1984). Teacher- versus peer-mediated instruction: An eco-behavioral analysis of achievement outcomes. *Journal of Applied Behavior Analysis,* **17,** 521–538.

Greenwood, C. R., Finney, R., Terry, B., & Arreaga-Mayer, C. (1990). *The classwide peer tutoring support program.* Kansas City, KS: Juniper Gardens Children's Project, Bureau of Child Research, University of Kansas.

Greenwood, C. R., Maheady, L., & Carta, J. J. (1991). Peer tutoring programs in the regular education classroom. In G. Stoner, M. R. Shinn, & H. M. Walker (Eds.), *Interventions for achievement and behavior problems* (pp. 179–200). Washington, DC: National Association for School Psychologists.

Greenwood, C. R., Terry, B., & Arreaga-Mayer, C. (1991, May). Using treatment fidelity data to examine treatment effects: Teacher's compliance with treatment schedules and effectiveness in the classwide peer tutoring program. In J. J. Carta (Chair), *Using treatment fidelity data to examine treatment effects.* Symposium presented at the 17th annual conference of the Association for Behavior Analysis, Atlanta, GA.

Greenwood, C. R., Terry, B., Arreaga-Mayer, C., & Finney, R. (1991, May). *Monitoring and maintaining fidelity of treatment using technology: The classwide peer tutoring expert system.* Group poster session presented at the 17th annual conference of the Association for Behavior Analysis, Atlanta, GA.

Harper, G., Mallette, B., Maheady, L., & Clifton, R. (1990). Applications of peer tutoring to arithmetic and spelling. *Direct Instruction News,* **9,** 34–38.

Hartmann, D. P., & Wood, D. D. (1982). Observational methods. In A. S. Bellack, M. Hersen, & A. E. Kazdin (Eds.), *International handbook of behavior modification and therapy* (pp. 109–138). New York: Plenum.

Jones, R. R., Vaught, R. S., & Weinrott, M. (1978). Effects of serial dependency on the agreement between visual and statistical inference. *Journal of Applied Behavior Analysis,* **11,** 277–283.

Kohler, F. W., & Greenwood, C. R. (1990). Effects of collateral peer supportive behaviors within the classwide peer tutoring program. *Journal of Applied Behavior Analysis,* **23,** 307–322.

Kohler, F. W., Richardson, T., Mina, C., Dinwiddie, G., & Greenwood, C. R. (1985). Establishing cooperative peer relations in the classroom. *The Pointer,* **29,** 12–16.

Maheady, L., & Harper, G. (1987). A classwide peer tutoring program to improve the spelling performance of low-income, third-, and fourth-grade students. *Education and Treatment of Children,* **10,** 120–133.

Maheady, L., Harper, G., Mallette, B., & Winstanley, N. (1991). *Training requirements in the use of classwide peer tutoring.* Manuscript submitted for publication.

Maheady, L., Sacca, M. K., & Harper, G. F. (1988). Classwide peer tutoring program on the academic performance of mildly handicapped students. *Exceptional Children,* **55,** 52–59.

O'Neill, R. E., Horner, R. H., Albin, R. W., Storey, K., & Sprague, J. R. (1990). *Functional analysis of problem behavior: A practical assessment guide.* Sycamore, IL: Sycamore.

Paine, S. C., & Bellamy, G. T. (1982). From innovation to standard practice: Developing and disseminating behavioral procedures. *The Behavior Analyst,* **5,** 29–44.

Peterson, L., Homer, A. L., & Wonderlich, S. A. (1982). The integrity of independent variables in behavior analysis. *Journal of Applied Behavior Analysis,* **15,** 477–492.

Schwartz, I. S., & Baer, D. M. (1991). Social validity assessments: Is current practice state of the art? *Journal of Applied Behavior Analysis,* **24,** 189–204.

Slavin, R. E. (1986). The Napa evaluation of Madeline Hunter's ITIP: Lessons learned. *Elementary School Journal,* **87,** 165–171.

Stallings, J. A., & Krasavage, E. M. (1986). Program implementation and student achievement in a four-year Madeline Hunter follow-through project. *Elementary School Journal,* **87,** 118–138.

Sulzer-Azaroff, B., & Gillat, A. (1990). Trends in behavior analysis in education. *Journal of Applied Behavior Analysis,* **23,** 491–495.

Terry, B., Greenwood, C. R., Arreaga-Mayer, C., Walker, D., & Finney, R. (1990). A computerized teacher advisor for evaluating student progress and solving implementation problems. In C. R. Greenwood (Chair), *Large scale dissemination of peer mediated intervention procedures: Issues, procedures, and results.* Symposium presented at the 16th annual convention of the Association for Behavior Analysis, Nashville, TN.

Yeaton, W. H., & Sechrest, L. (1981). Critical dimensions in the choice and maintenance of successful treatments: Strength, integrity, and effectiveness. *Journal of Consulting and Clinical Psychology,* **49,** 156–167.

Received September 9, 1991
Initial editorial decision November 8, 1991
Revisions received November 20, 1991; November 21, 1991
Final acceptance December 3, 1991
Action Editor, Susan A. Fowler